Using a parser as a heuristic tool for the description of New Englishes

Schneider, Gerold ; Hundt, Marianne

Abstract: We propose a novel use of an automatic parser as a tool for descriptive linguistics. This use has two advantages. First, quantitative data on very large amounts of texts are available instantly, a process which would take years of work with manual annotation. Second, it allows variational linguists to use a partly corpus-driven approach, where results emerge from the data. The disadvantage of the parser-based approach is that the level of precision and recall is lower. We give evaluations of precision and recall of the parser we use. We then show the application of the parser-based approach to a selection of New Englishes. using several subparts of the International Corpus of English (ICE). We employ two methods to discover potential features of New Englishes: (a) by exploring quantitative differences in the use of established syntactic patterns (b) by evaluating the potential correlation between parsing break-downs and regional syntactic innovations.

Using a parser as a heuristic tool for the description of New Englishes

Gerold Schneider and Marianne Hundt
English Department, University of Zurich
{gschneid,mhundt}@es.uzh.ch

ABSTRACT

We propose a novel use of an automatic parser as a tool for descriptive linguistics. This use has two advantages. First, quantitative data on very large amounts of texts are available instantly, a process which would take years of work with manual annotation. Second, it allows variational linguists to use a partly corpus-driven approach, where results emerge from the data. The disadvantage of the parser-based approach is that the level of precision and recall is lower. We give evaluations of precision and recall of the parser we use.
We then show the application of the parser-based approach to a selection of New Englishes. using several subparts of the International Corpus of English (ICE). We employ two methods to discover potential features of New Englishes: (a) by exploring quantitative differences in the use of established syntactic patterns (b) by evaluating the potential correlation between parsing break-downs and regional syntactic innovations.

1. Introduction

The detailed description of varieties like Philippine (PhilE) or Singapore English (SingE) is still a relatively young field in English linguistics. Furthermore, the grammatical description initially relied largely on anecdotal evidence, (see Foley 1988, Bautista and Gonzales, 2006) or at best on the (mainly manual) analysis of sociolinguistic fieldwork projects (see Schreier 2003 on a lesser-known variety of English) rather than large, representative collections of texts. Standard reference corpora of these varieties, i.e. the different regional components of the *International Corpus of English*, have become available relatively recently: ICE-Philippines was released in 2005, but ICE-Jamaica only came out this year, for instance. So far, the exploitation of these new sources of data has been in verifying previously existing hypotheses on regionalisms or the systematic comparison of these new Englishes (often with British and/or American English as the yardstick of comparison). Furthermore, the corpus-based descriptions that are becoming available (see Sand 1999, Schneider 2004 or Sedlatschek 2009) are based on orthographic corpora which have not been annotated, as work on tagging and parsing the ICE components is still under way.[i] As a result, these descriptions have to rely on more or less sophisticated searches based on lexical items. Our idea is that it might be worthwhile investigating whether it is possible to arrive at a partly corpus-driven description of the New Englishes. In this approach, the available corpora are annotated grammatically. The aim is to explore in how far this annotation process in itself might yield useful information for the description of the New Englishes and that these, in turn, might be exploited in fine-tuning the annotation tools to the structural challenges that New Englishes present. Initially, we are using existing ICE corpora for our pilot study, but the aim is to apply the same methodology to larger, web-derived (and thus somewhat 'messier') data.

In part two of our paper, we will give a brief overview of the corpora that we used and comment on the annotation of the corpus material. In part three of the paper, we will outline our methodology which basically explores two corpus-driven approaches: one that starts from quantitative differences between the annotated data, whereas the other one relies more heavily

on the process of evaluating the automatically annotated corpora. In section 4, the overall evaluation of the parser output will be briefly evaluated. The results of our study are given in part 5 of our paper.

2. Data

We compare ICE-Fiji (and partly ICE-India) as L2 varieties to ICE-GB and (and partly to ICE-NZ) as L1. We have used all finished subparts of the written ICE-FIJI corpus where writer ethnicity and gender is known, about 250,000 of the planned 400,000 words. The selection of subtexts and the broad genre label is summarised in the following table.

Text	Used	Unused	Genre
W1A	20	0	Essay
W1B	0	20	
W2A	40	0	Academic
W2B	25	15	Non-Academic
W2C	20	0	Press
W2D	0	20	
W2E	9	1	Press
W2F	0	20	
TOTAL	114	76	

Table 1: Composition of sub-corpora

For our comparison to other ICE corpora, we have only used the texts whose corresponding text in ICE-FIJI is used in our selection.

3. Methodology

The background of our methodology is summarised in Figure 1. So far, corpus analyses were informed by previous research. We use computational tools to help in the annotation process. The errors that the tools make are expected to feed back into improvements of the computational tools. The results based on more richly annotated data will feed into the description and theory of the new Englishes. This incremental cycle is illustrated in the orange box.

As pointed out above, we are using a two-pronged approach in our investigation:

(a) We assume that the parser has correctly annotated the sentences in our corpora and that statistical differences in the use of certain constructions accurately represent differences between varieties. This is a relatively traditional approach (see Hundt 1998) but the difference is that we are making use of heavily annotated data rather than the comparison of word frequencies.

(b) New Englishes are characterized by structural nativization, i.e. the establishment of features that range from very localized patterns (e.g. the *kena*-passive or the *already*-perfect in SingE, see (1) and (2), respectively) to more widely attested patterns (e.g. uninflected participles, copula absence, zero determiners and divergent agreement patterns, as illustrated in examples (3)-(6)). These are expected to be potential breaking points in the parsing process.

1) His tail like like *kena caught* in the in the ratch hut (ICE-SING S1A-052)

2) a. A lot of people *finish already*. (ICE-SING S1A-069)

   b. … we all *makan* [eat] *already* (ICE-SING S1A-023)

3) So they have ask for extension five times already (SING S1A-051)

4) Thus, *more employment available* for women and men. (ICE-FIJI W1A-018)

5) The other very obvious omission is Ø exclusion of some of the peri-urban schools whose rolls have been expanding rapidly and are in great need of financial assistance to take care of their expanding school roll. (ICE-FIJI W2C-013)

6) a. Women play*s* a vital role in mostly all the societies … (ICE-FIJI W1A-020)

b. Chaudry distort facts on sugar deal … (ICE-FIJI W2C-006)

If and in how far these (and hitherto undetected) nativized patterns can be uncovered in a partly corpus-driven approach will be explored in section 5.2.

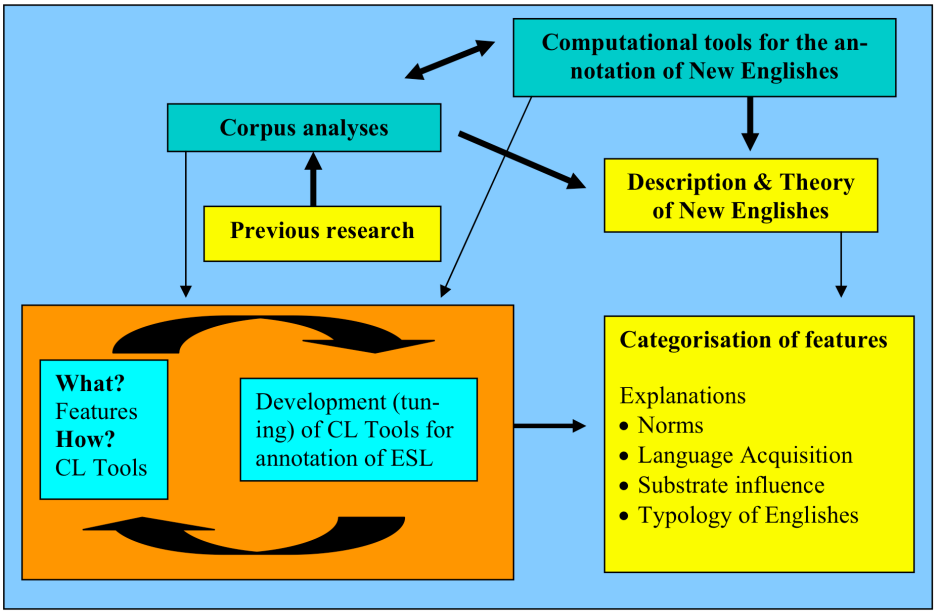

Figure 1. Visualisation of the process

3.1 Profiting from grammatical Annotation

Corpora that are syntactically annotated allow linguists to conduct new types of research. Smadja (1993), for example, expresses the desire for tools that annotate large texts that were not available in the early 1990s.

> Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure. However, in practice, freestyle texts contain a great deal of nonstandard features over which automatic parsers would fail. This fact is being seriously challenged by current research [...] and might not be true in the near future. (Smadja 1993: 151)

Recent progress in parsing technology (e.g. Collins 1999, Schneider 2008) has now made it possible to parse large amounts of data. But automatic annotation incurs a certain level of errors, so manually (or part-manually) annotated texts would be preferable. They are available, for example ICE-GB[ii] and the Penn Treebank. These corpora are valuable resources, but they are too small for some types of research (for example collocations) and they cannot be used for variationist linguistics until other ICE corpora than ICE-GB are provided with the same kind of syntactic annotation. For such types of research, using automatically parsed texts is a viable alternative, even when taking parsing errors into account. Despite Smadja's call for syntactically annotated data, the use of parsers in corpus linguistics is still quite rare, but some initial (and successful) applications have been reported, for example:

- Syntax-Lexis interaction and collocation (e.g. Seretan and Wehrli 2006, Lehmann and Schneider 2009)
- Psycholinguistics (e.g. Schneider et al. 2005)

- Language variation (e.g. ongoing work at UCL by Bas Aarts and Joanne Close[iii] on diachronic variation in the English Verb Phrase and ongoing research by Hundt and Schneider, see Hundt and Schneider fc.)

3.2 The grammar as model

The parser we use – Pro3Gres – combines a competence grammar model with a performance disambiguation model. The grammar has been written manually and carefully designed to cover the most frequent phenomena of standard L1 English grammar. As such, it can serve as a model of such varieties of English. In a first step, one could assume that such a model can be used for testing grammaticality: sentences that are fully parsed are grammatical, whereas sentences that contain marked deviations from the model will fail to parse. Note that such parsing fails might also occur in texts produced by speakers of L1 varieties, as we will see later on.[iv]
The parser has been designed to strike a balance between speed and accuracy for Information Retrieval applications. It is designed to be as robust as possible, fast for application over very large corpora,[v] and to keep search spaces manageable by excluding some rare and very marked structures. This complicates our initial assumption in the following way.

- Making the parser robust means that it is able to deal with any real world output. Many of the sentences occurring in large corpora do not closely adhere to (school) grammar models, so using strict grammar models is largely seen as a failed attempt in modern parsing technology. For example, the parser does not strictly enforce agreement, and it uses statistical preferences instead of strict sub-categorisation frames and selectional restrictions; this entails that e.g. prepositional phrases with divergent prepositions get attached. The parser has been applied to several hundred million words, among others the BNC, without crashing. If sentence complexity gets high, or if marked garden path situations arise, or if the parser does not find a suitable grammar rule, it often reports several fragments for a sentence. We will show in section 5.2 that robustness counteracts our initial goal of using the parser as a heuristic tool to a considerable extent.
- In order to keep search spaces manageable and error rates low, local ambiguity levels need to be kept low, and very rare phenomena are not included in the grammar. For example, violations of X-bar constraints and strong heaviness ordering violations are not licensed by the grammar.

7) He confirmed to Mary that he will go.

8) He confirmed that he will go to Mary.

Sentence 8) is ruled out to be analysed as having the same interpretation as sentence 7). The PP *to Mary* in 8) can only attach to the verb *go*, it is not allowed to cross the heavy clause *that he will go*. This means that not only nativized patterns of New Englishes will break the parse but also rare but grammatically accepted phenomena in our reference variety, BrE.

4. Evaluation

Automatic parsing leads to results that contain a certain level of errors. In the following, we assess the error level. To this end, we use an evaluation corpus of standard British English (BrE) to assess the performance of Pro3Gres. In a second step, we manually evaluate the performance of Pro3Gres on small random samples of ICE-GB and on ICE-FIJI.
As our evaluation corpus of standard British English, we use GREVAL (Carroll et al. 2003), a 500 sentence random excerpt from the Susanne corpus that has been manually annotated for syntactic dependency relations. Performance on subject, object and PP-attachment relations is

given in Table 2. More details on the evaluation can be found in Schneider (2008). Some of the errors are due to grammar assumptions that vary between the annotation scheme used in GREVAL and the parser output.

Peformance on GREVAL	Subject	Object	Noun-PP	Verb-PP
Precision	92%	89%	74%	72%
Recall	81%	84%	66%	84%

Table 2: Performance of Pro3Gres on the 500 GREVAL sentences

The statistical component of the parser has been trained on the Penn Treebank that contains a genre mix which is slightly different from the ICE corpora. Differences across registers, however, may be less problematic for parsing than marked differences across regional varieties of English. Especially L2 corpora are expected to contain regional patterns that are considerably different from L1 data. To evaluate the parser by comparing its performance on L1 English and an L2 variety is very important, for at least two reasons:

- Prong (a): parsing performance on different varieties of English may be considerably lower. It is well known that using parsers on genres that are slightly different leads to lower performance. If performance on an L2 variety, for example Fiji English (FE), is considerably lower than on a corpus of BrE, results obtained by method (a) are seriously called into question.
- Prong (b): the evaluation will show if the parser produces consistent errors or break-downs on constructions that are typical of L2 Englishes, and may thus help to accept or refute method (b), i.e. the use of the parsing tool as a heuristic in the description of new Englishes.

In order to assess the performance of the parser on ICE-GB and on ICE-FIJI, we manually checked the output of 100 random sentences from each. Since this evaluation method (a posteriori checking) is slightly less strict than a priori annotation, values are generally higher than on GREVAL, to which they cannot be directly compared. Between ICE-GB and ICE-FIJI, performance can be compared, since the evaluation method was identical. The results on parser performance are given in Table 3.

ICE-GB	Subject	Object	Noun-PP	Verb-PP
Precision	58/60 = 97%	45/49 = 92%	23/28 = 82%	35/39 = 90%
Recall	58/63 = 92%	45/51 = 88%	23/29 = 79%	35/38 = 92%

ICE-FIJI	Subject	Object	Noun-PP	Verb-PP
Precision	71/72 = 99%	44/47= 94%	43/51 = 84%	45/58 = 72%
Recall	71/73 = 97%	44/44= 100%	43/47 = 91%	45/59 = 76%

Table 3: A Posteriori Checking Performance on ICE-GB and ICE-FIJI

Table 3 reveals that the general performance of the parser on the two types of corpora is similar. Verb PP-attachment performance on ICE-FIJI is considerably lower, whereas performance on the subject and object relations turns out to be slightly higher. The result on subject and object relations may be affected by the fact that sentences in ICE-FIJI are shorter than in ICE-GB. Counts are quite low, so that fluctuation is quite high, the 100% recall on ICE-FIJI *object* is probably due to chance.

The fact that performance on the verb PP-attachment relations is lower on ICE-FIJI than on the British reference corpus may partly be related to some of the constructions found in FE. For example, in sentence 9) in our evaluation random set, the parser attaches *from the world* to

rate instead of to *demolish* (see **Error! Reference source not found.**), because verbal attachment of a PP introduced by *from* to *demolish* (a frame which does not exist in standard English) is judged even less likely than nominal attachment of this PP to the noun *rate* (in financial texts, which are frequent in the Penn Treebank, rates often start at a value indicated by *from,* so the probability that a PP introduced by *from* attaches to *rate* is quite high). The recall error arising from this statistical model means that not all instances showing such constructions will be found. But instances appearing in a less ambiguous context will be found, so that even such errors may not have to discredit our prong (a).

9) […] in order to demolish the poverty rate from the world, women should keep on fighting for their rights. (ICE-FIJI W1A-017)

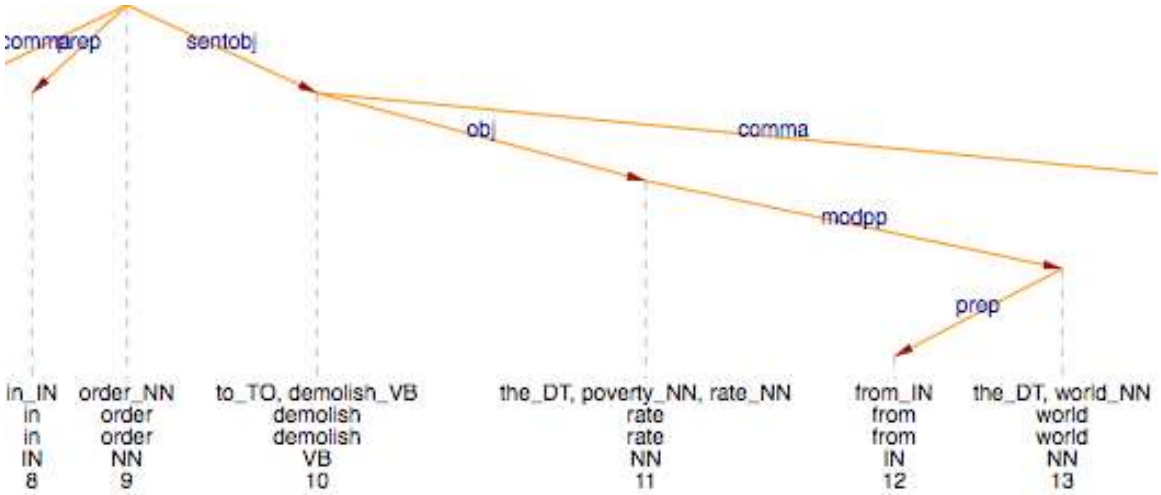

Figure 2. Syntactic analysis of relevant part of sentence from ICE-FIJI W1A-017

In conclusion, it can be said that performance on substantially different texts is not dramatically lower, and that some of the ESL structures lead to a slightly higher error level, or in other words: many, but not all instances of frequency-based ESL features will be found using prong (a). More importantly, no pattern of consistent errors emerged from our small evaluation; this and the fact that performance on L2 corpora is generally similar to performance on L1 data do not point in favour of prong (b).

We pointed out above that, in those cases where the parser fails to find a suitable grammar rule, it often reports several fragments for a sentence. Another potential avenue for uncovering previously undetected patterns of nativization (i.e. following prong (b))may therefore be in the systematic analysis of fragmented parser output. We will do so in section 5.2.

5. Results

In our pilot study, we focus on an example each for the two approaches in which the parser might be used as a heuristic. In section 5.1, we look at relational profiles, i.e the frequencies with which certain syntactic categories are used across our corpora. Obviously, in a further step, any statistically significant differences will be the starting point for further qualitative analyses. In section 5.2, we investigate the possibility of exploiting parsing errors and parser break-downs as a discovery procedure for previously undetected features of ESL varieties by looking at (a) the fragmentation of analyses and (b) probability scores.

5.1 Relation profiles

In this section, we report on the comparison of the frequencies of some important relation types that the parser delivers. Apart from the obvious subject and object relation, we include noun-PP attachment (*modpp*), verb-PP attachment (*pobj*), possessives (*pos*), ditransitives (*obj2*), nouns that are postmodified by relative clauses (*modrel* and *modpart*). Table 4 gives an overview of the results. The relation labels have the following meaning: *subj* for subject, *obj* for object, *modpp* for PP-attachment to noun, *pobj* for PP-attachment to verb, *pos* for possessive (also called saxon genitive), *obj2* for secondary objects, *modrel* for relative clauses, *modpart* for reduced relative clauses, *detmod* for the noun-determiner relation.

Relation	ICE-FIJI		ICE-INDIA		ICE-GB		ICE-NZ	
per sentence	Occurrence of Relation	Per sentence	Occurrence of Relation	Per sentence	Occurrence of Relation	Per sentence	Occurrence of Relation	Per sentence
subj	16112	1.72	16189	1.45	17334	1.70	19357	1.71
obj	9870	1.05	10145	0.91	10761	1.06	11865	1.04
modpp	10456	1.12	13276	1.19	12464	1.22	13916	1.22
pobj	10491	1.12	10911	0.98	12463	1.22	13141	1.16
pos	739	0.07	718	0.06	746	0.07	1052	0.09
obj2	53	0.01	70	0.01	77	0.01	92	0.01
modrel	1522	0.16	1463	0.13	1834	0.18	1800	0.16
modpart	1030	0.11	1059	0.09	1032	0.10	1466	0.13
detmod	21456	2.30	23755	2.13	24632	2.42	27610	2.44

Table 4: Frequency of Relations in ICE-INDIA, ICE-FIJI, ICE-GB, and ICE-NZ, across genres[vi]

The differences are statistically significant according to the chi-square contingency table test ($p < 0.001$) In the per sentence count, some of the differences are due to sentence length. Sentence length in ICE-FIJI is 21.8 words on average, in ICE-INDIA it is 20.4, in ICE-GB it is 22.5 words on average, in ICE-NZ it is 23.3.[vii] In order to normalise for sentence length, we report figures per 1000 words in Table 5.

Relation	ICE-FIJI		ICE-INDIA		ICE-GB		ICE-NZ	
per 1000w	Occurrence of Relation	Per 1000w	Occurrence of Relation	Per 1000w	Occurrence of Relation	Per 1000w	Occurrence of Relation	Per 1000w
subj	16112	78.9	16189	71.1	17334	75.6	19357	73.4
obj	9870	48.2	10145	44.6	10761	47.1	11865	44.6
modpp	10456	51.4	13276	58.3	12464	54.2	13916	52.4
pobj	10491	51.4	10911	48.0	12463	54.2	13141	49.8
pos	739	3.2	718	2.9	746	3.1	1052	3.9
obj2	53	0.5	70	0.5	77	0.4	92	0.4
modrel	1522	7.3	1463	6.4	1834	8.0	1800	6.9
modpart	1030	5.0	1059	4.4	1032	4.4	1466	5.6
detmod	21456	105.5	23755	104.4	24632	107.6	27610	104.7

Table 5: Frequency of Relations in ICE-INDIA, ICE-FIJI, ICE-GB, and ICE-NZ, across genres, normalised by sentence length

We observe that the *pobj* relation (a PP attached to a verb) is rarer in ICE-INDIA than in the other ICE corpora. We further observe that the *modpp* relation (a PP attached to a noun) is considerably rarer in ICE-FIJI. It is very frequent in ICE-INDIA if we consider sentence length. One explanation for the high frequency of *modpp* in Indian English (IndE) is that it is dominated by PPs introduced by *of*: in ICE-INDIA, 53.4% of all *modpp* relations are headed by the preposition *of*. The corresponding percentage in ICE-GB is 43.7%, in ICE-FIJI it is 50.3%, in ICE-NZ it is 49.8%.

We have broken down the *modpp* relation by genre (Table 6). Genre differences are very marked: academic texts contain complex NPs in all varieties, essays considerably less, especially in ICE-INDIA.

modpp per sentence	ICE-FIJI Occurrence of Relation	Per sentence	ICE-INDIA Occurrence of Relation	Per sentence	ICE-GB Occurrence of Relation	Per sentence	ICE-NZ Occurrence of Relation	Per sentence
essay	2152	1.09	2157	0.78	2547	1.17	2359	1.08
academic	3048	1.39	4693	1.47	4294	1.54	5808	1.42
non-academic	2322	1.12	2667	1.21	2317	1.09	2624	1.15
press	2934	0.94	3759	1.24	3306	1.07	3125	1.13

Table 6: Frequency of noun-modifying PPs in ICE-INDIA, ICE-FIJI, ICE-GB and ICE-NZ by genre

Another interesting finding that emerges from Tables 4 and 5 is that secondary objects are used less frequently in ICE-INDIA and ICE-FIJI than we would have expected. Previous research (Mukherjee 2005) indicates that a more frequent use of verbs in ditransitive constructions is typical of IndE; we suspect that it might also be a feature of FE (due to the presence of speakers with an Indian background). We have therefore broken down ditransitive instances by head verb (Table 7).

ICE-FIJI	ICE-INDIA	ICE-GB	ICE-NZ
33 give	32 give	35 give	55 give
5 call	9 provide	6 offer	6 do
4 tell	6 offer	4 pay	5 tell
3 provide	2 show	4 cost	5 call
2 show	2 promise	4 call	4 offer
2 pay	2 grant	3 do	4 find
2 leave	2 do	2 tell	3 show
2 allow	2 develop	2 show	3 leave
1 teach	2 ask	2 hand	3 allow
1 stop	1 tell	2 grant	2 serve
1 save	1 teach	2 earn	2 save
1 refuse	1 serve	2 bring	2 name
1 grant	1 save	2 allow	2 cost
1 find	1 pay	1 send	2 consider
1 deny	1 lend	1 provide	1 win
1 ask	1 hand	1 permit	1 pay
1 accord	1 deny	1 name	1 lend
	1 consider	1 leave	1 earn
	1 call	1 award	1 deny
	1 bring	1 ask	1 bring
			1 ask

Table 7: Ditransitive constructions (head verbs) in ICE-FIJI, ICE-INDIA, ICE-GB and ICE-NZ

Although data is sparse, this break-down shows that *offer* is rare in ICE-FIJI,[viii] and in ICE-INDIA. Ditransitive use of *provide*, on the other hand, which is one of the constructions described by Mukherjee (2005) as a ditransitive typical of IndE, is far more frequent in ICE-FIJI and ICE-INDIA. Furthermore, a close look at the parses reveals that the one ditransitive use of *provide* in ICE-GB is actually a parser error,[ix] while the three instances in ICE-FIJI and the nine occurrences in ICE-INDIA are correctly parsed. An example is given in Figure 3.

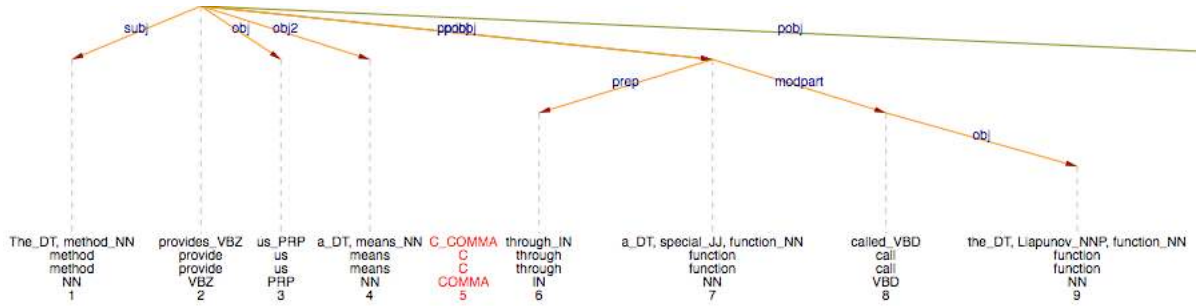

Figure 3: Parse of a sentence from ICE-FIJI (W2A-030): "The method provides us a means, through a special function called the Liapunov function, of looking at the whole class of solutions; for example, those that converge to the equilibrium state (asymptotic stability) and those that diverge and leave the defined boundary of stability (instability)."

Finally, verbs like *call* and *name* appear in Table 7 because the parser treats *elect* verbs as ditransitives.

The break-down of ditransitive constructions in our corpora is an example of a more qualitative analysis. It shows that statistical findings which, at first, appear to be at odds with previous research may be partly substantiated on the level of qualitative analysis.
It is often necessary to consider fine-grained distinctions, be it by investigating the data by genre as we did for noun-modifying PPs, or by lexical head, as we did for ditransitive verbs. Often, only more fine-grained analyses reveal differences. The determiner relation *detmod*, for example, seems to be similarly frequent across all corpora (see Table 5). The analysis per genre, however, reveals surprising differences (see Figure 4), but no clear picture emerges, neither on a per sentence nor on a per 1000 words count, because both sentence length and noun complexity vary considerably across the genres. If one measures the counts per noun, however, a clear tendency emerges: determiners in less edited genres are considerably rarer in ICE-INDIA and ICE-FIJI than in ICE-GB and ICE-NZ. If we count determinerless nouns that could have a determiner, namely singular non-proper nouns, we get counts for zero-determiners. The distribution of zero-determiners is shown in Figure 4. The percentage shows how many singular non-proper nouns lack a determiner. ICE-INDIA shows a clear preference for zero-determiners in the essay genre and partly in academic writing, but it also reveals hypercorrection in the highly edited press genre. ICE-FIJI shows similar tendencies, although less strongly.

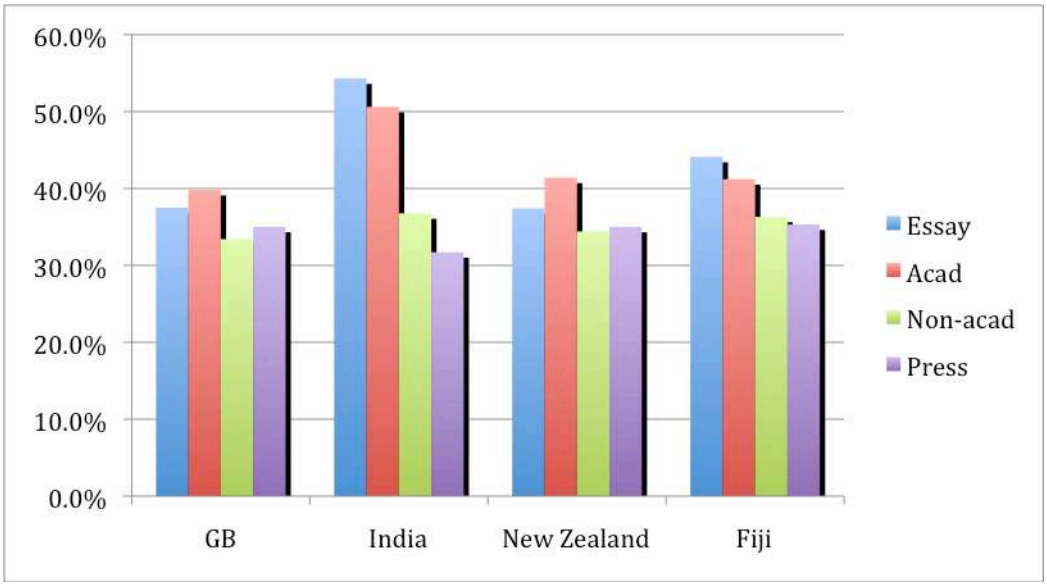

Figure 4: The distribution of zero-determiners across different varieties and genres.

5.2 Exploiting Parsing Errors and Break-Downs

We mentioned that, in an initial step, we assume that the grammar could be used for testing grammaticality, (approach (b) described in section 3), but that the robustness, speed and relative accuracy of the parser partly oppose this assumption. On the one hand, there are very many perfectly acceptable sentences where the parser reports several fragments instead of one complete span (i.e. the parser breaks down on grammatical sentences); on the other hand, there are many divergent sentences where the parser reports a wrong or questionable analysis that spans the entire sentence.

5.2.1 Fragmentation of Analyses

In this subsection we investigate whether there is a correlation between acceptability and the number of parser break-downs, where a break-down is defined as the event where the parser reports several fragmentary analyses. We will use the term fragmentation. A fragmentation value of 1 means that the parser assigns a complete analysis to every sentence; in other words, the parsing of the sentence is not actually fragmented. A fragmentation value of 2 accordingly means that the parser assigns 2 fragmentary analyses per sentence on average. Fragmentation is shown in table 8.

	ICE-FIJI	ICE-INDIA	ICE-GB	ICE-NZ
Average number of fragments	1.58	1.59	1.68	1.76
# fragments	14812	17827	17207	19987
# sentences	9331	11152	10190	11315
Fragments per word	7.2%	7.8%	7.5%	7.6%

Table 8: Fragmentation in ICE-FIJI, ICE-INDIA, ICE-GB and ICE-NZ

Differences are quite small but obvious. How are we to interpret them? The lower number of fragments in ICE-FIJI and ICE-INDIA largely reflects, again, that sentence length and complexity is lower than in ICE-GB or ICE-NZ. Longer sentences, as a rule, are considerably more likely to cause fragmentation than short ones. The last line in the table shows the number of fragments per words, which is the average number of fragments divided by the average number of words. This number indicates the chance of a word to 'break the parse', and shows very little difference.

Genre Variation

Features of L2-Englishes are considerably more noticeable in less edited genres. In other words, printed texts will contain fewer instances of nativized patterns than unprinted written texts; these, in turn, will contain fewer nativized constructions than spontaneous speech. We have, for example, shown that zero articles are used significantly more often in the category 'essay' in ICE-India and ICE-FIJI than in ICE-GB (see Hundt and Schneider, fc.) In the carefully edited 'press' genre, zero articles are similarly infrequent in all ICE corpora, and there is even some hypercorrection in ICE India. Zero articles per base noun chunk are given in Table 9.

Zero Articles by Genre, per noun	ICE-FIJI	ICE-INDIA	ICE-GB	ICE-NZ
Essay	0.489	0.592	0.428	0.445
Acad	0.459	0.546	0.411	0.444
Non-acad	0.431	0.427	0.398	0.394

Press	0.410	0.376	0.406	0.398
∑	0.443	0.49	0.41	0.425

Table 9: Zero Determiners by genre in ICE-FIJI, ICE-INDIA, ICE-GB and ICE-NZ

If this assumption holds, then fragmentation in the 'essay' genre might be higher. Table 10 summarizes the results on fragmentation by genre.

Fragmentation	ICE-FIJI			ICE-GB		
by genre	Sentences	Fragments	Ratio	Sentences	Fragments	Ratio
Essay	1960	3338	1.70	2181	3606	1.65
Academic	2187	3485	1.59	2786	4914	1.76
Non-Academic	2063	3277	1.59	2124	3581	1.69
Press	3121	4712	1.51	3099	5106	1.64
TOTAL	9331	14812	1.58	10190	17207	1.68

Table 10: Fragmentation by genre in ICE-FIJI and ICE-GB

Table 10 shows that genre differences are small and quite similar between ICE-Fiji and ICE GB, with only two marked differences: Fragmentation in the Fiji essays and in British academic writing is particularly high. Again, sentence length and complexity is a crucial factor that these figures do not reflect yet. Sentence length and complexity is considerably lower in the 'essay' genre, and considerably higher in the 'academic' genre (see Table 11).

Sentence length	ICE Fiji			ICE GB		
by genre	Sentences	Words	Sent length	Sentences	Words	Sent length
Essay	1960	43719	22.3	2181	46864	21.5
Academic	2187	50187	22.9	2786	67765	24.3
Non-Academic	2063	44222	21.4	2124	47004	22.1
Press	3121	65398	21.0	3099	67736	21.9
TOTAL	9331	203526	21.8	10190	229369	22.5

Table 11: Sentence length by genre in ICE-FIJI and ICE-GB

What is perhaps surprising is that in the academic texts in ICE-FIJI, sentences are not particularly long, whereas the 'essay' section yields long sentences. Table 11 also clearly shows that the high fragmentation in the 'academic' subsection of ICE-GB is largely related to its high complexity: sentences are 1.8 words longer than on average, and fragmentation is 0.08 higher. In ICE-FIJI, on the other hand, 'essay' sentence length is 0.5 words longer than average, and fragmentation is 0.12 higher.

As fragmentation and sentence length are not necessarily linearly related, normalising fragmentation by sentence complexity may shed additional light on this comparison. The percentage in Table 12 represents the chance for a word to 'break the parse'.

Fragmentation	ICE Fiji			ICE GB		
by sentence length	Fragmentation	Sent.Length	Ratio	Fragmentation	Sent.Length	Ratio
Essay	1.70	22.3	7.62%	1.65	21.5	7.67%
Academic	1.59	22.9	6.94%	1.76	24.3	7.24%
Non-Academic	1.59	21.4	7.43%	1.69	22.1	7.65%
Press	1.51	21.0	7.19%	1.64	21.9	7.49%
TOTAL	1.58	21.8	7.25%	1.68	22.5	7.47%

Table 12: Ratio of fragmentation and sentence length by genre in ICE-FIJI and ICE-GB

Table 12 shows that differences are relatively small. The figures reflect that fragmentation is generally a little higher in unedited styles, both in ICE-FIJI and ICE-GB.

Correlation between Break-Downs and anecdotal observation

During the corpus compilation process, ICE Fiji was unsystematically tagged for features that might potentially be typical of FE. About 10% of the sentences in ICE-FIJI were thus annotated with the tag, <OBS!>. While this markup was unsystematic, i.e. yielding incomplete recall, its precision can be assumed near-perfect. We tested whether words marked with an <OBS!> tag are likely to break the parse. For this, we extracted 20 random instances of <OBS!> tags and checked the parser output. Surprisingly, there are only three cases where the marked phenomenon breaks the parse. (15%, compared to a chunk's base chance of 7%). Although our random sample is small, this test again indicates that there is only very little connection between potential features of FE and fragmentation (i.e. parser failure).

In a second step, we grouped the 20 random instances into classes in order to assess frequent types of FE features. We found the following classes:

- divergent agreement (see sentence 6) at the beginning of chapter 3)
- unusual tense and aspect marking (see sentence 3) in chapter 3)
- lacking or hypercorrected determiner (see sentence 5) in chapter 3)
- unusual collocations and idioms

There were five instances of divergent agreement, for example:

10) Women <OBS!>plays</OBS!> a vital role in mostly all the societies, but it is how they are looked upon in the society at large. (ICE-FIJI, W1A-020, sentence 2317)

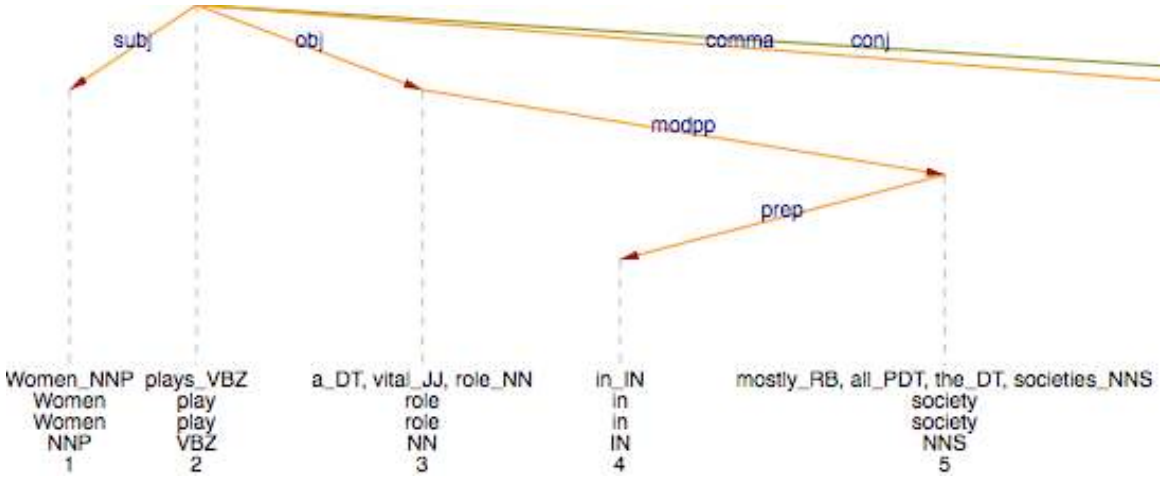

Figure 5: Syntactic analysis of relevant part of ICE-FIJI, W1A-020, sentence 2317

There were four instances of unusual tense and aspect marking, for example:

11) The Asia Pacific region has been described as <OBS!>been </OBS!> both 'volatile and dynamic'. (ICE-FIJI, W1A-005, sentence 496)

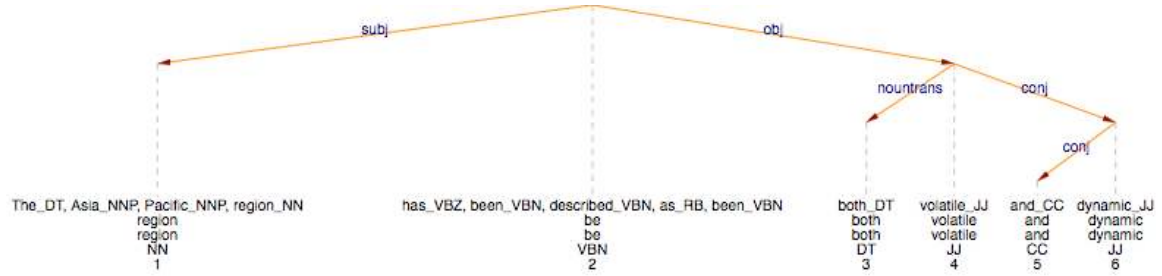

Figure 6: Syntactic analysis of ICE-FIJI, W1A-005, sentence 496

There were four instances of lacking (or hypercorrected) determiner, for example:

12) Since the NCLChs was not crosslinked , the polymer chains were flexible and <OBS!>increase</OBS!> in temperature caused breaking of secondary interactions, creating more space for water within the matrix of the gel. (ICE-FIJI, W2A-027, sentence 4335)

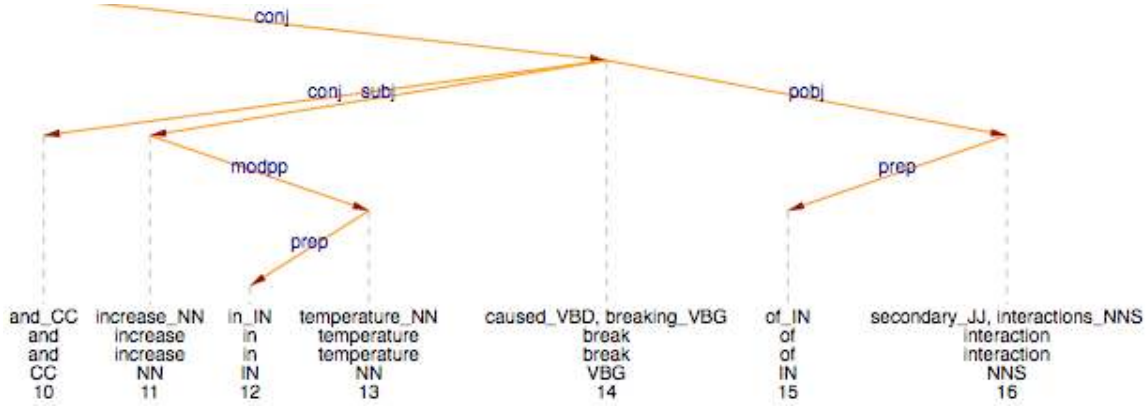

Figure 7: Analysis of the relevant part of ICE-FIJI, W2A-027, sentence 4335

There were three instances of unusual collocations and idioms, for example:

13) She had been beautiful when she was younger and her parents thought that out of all their children Sylvia would <OBS!>do</OBS!> them proud by marrying someone with more money than them . (ICE-FIJI, W2F-009, sentence 12740)

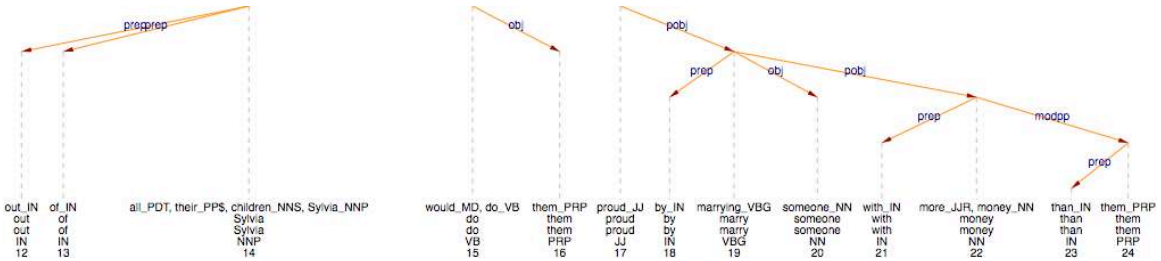

Figure 8: Syntactic analysis of ICE-FIJI, W2F-009, sentence 12740

There was one case of anaphor agreement violation, one case of an unusual *s*-genitive (sentence 14)), and two cases that we could not clearly classify.

14) Therefore , the governments should consider the views that <OBS!>women's</OBS!> present in women's workshops and seminars and take necessary means and ways that there is equality between men and women and they can work hand in hand to have a peaceful, respectable and harmonious family, society, country and nation as a whole. (ICE-FIJI W1A-020)

The three cases where the marked phenomenon breaks the parse are spread quite homogenously across the Fijian feature classes just introduced. There is one unusual use of tense, one unusual collocation (sentence 13), Figure 8), and one divergent agreement structure. Thus, no clear pattern emerges, and the parser proves surprisingly robust – which was one of its design goals.

In the majority of the cases where the parse is not fragmented it is actually correct (see sentences 10) to 12)). The error rate seems to be higher, though, than on sentences that do not show FE features, but we will need larger amounts of data to assess this.

Some of the examples of structural nativization given at the beginning of section 3 are absent (as we expect for sentences 1) and 2)) or not very frequent (sentence 4)) in ICE-FIJI; no

instances of them appear in our 20 random instances. The parser reports the following syntactic analyses for them (Figures 9-11). Again, the parser shows robust behaviour.

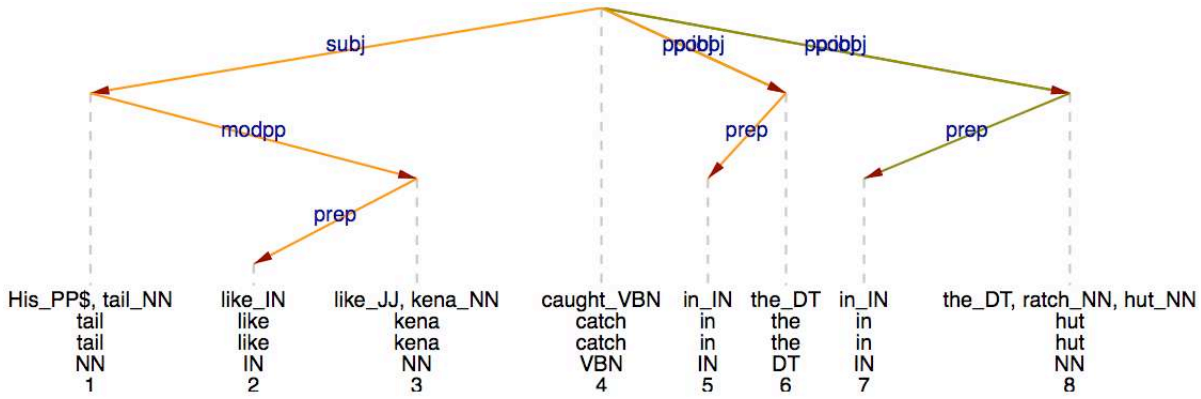

Figure 9: Parse of sentence 1)

The unknown word *kena* in Figure 9 is given a noun tag, triggered by the fact that many words following an adjective are nouns.

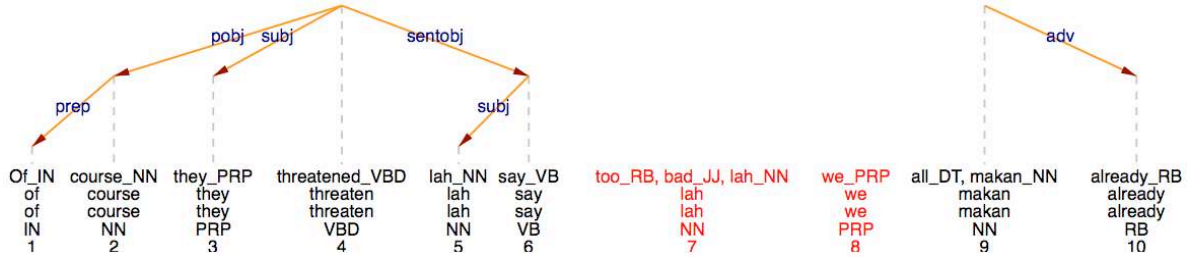

Figure 10: Parse of sentence 2)b

The unknown word *makan* in Figure 10 is given a noun tag, triggered by the fact that many words following a determiner are nouns. The parse breaks, but at a different point.

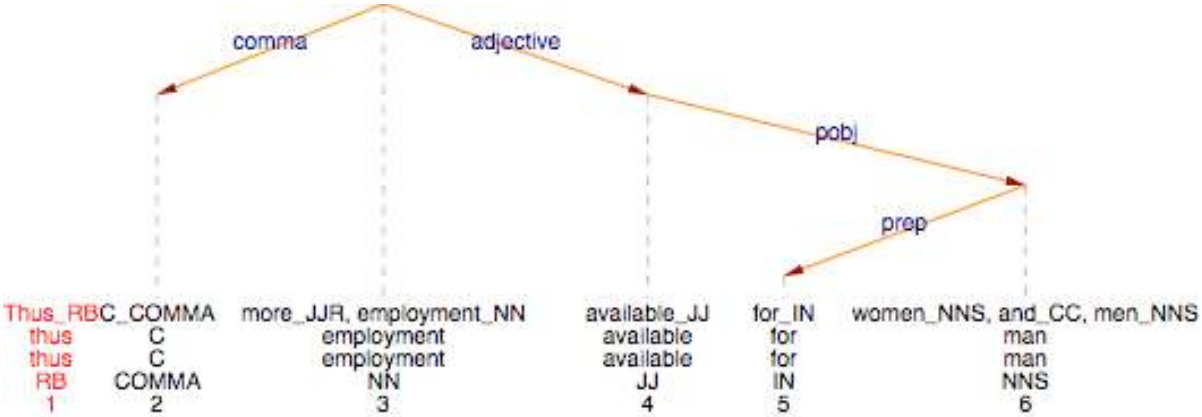

Figure 11: Parse of sentence 4)

The *adjective* relation typically expresses postmodification by adjective, as in *productive* ***resources sufficient*** *to ensure sustainable livelihood* (ICE-FIJI W1A-018). In a deep-syntactic analysis, the *adjective* relation is akin to a reduced relative and often mapped to subject function, thus isomorphic to the intended analysis.

In one sense, then, the robustness of the parser output is bad news because we do not seem to be able to exploit incorrect parser output as a discovery procedure for hitherto undetected localised grammar in New Englishes. In another sense, this is good news in that one can specifically search for the potential FE features outlined above (or more generally for L2 feature classes that one has observed), thus following prong (a) in our approach, as outlined in section 3. We will report on searches for the potential FE features in a separate paper.

5.2.2 Probability Scores

The probabilistic performance disambiguator, which is trained over the syntactically annotated Penn Treebank, can also be used as a model (see section 3.2 above). On the one hand, the probability model is used for ranking the very many possible parses of a sentence, increasing the probability that the correct analysis will be reported by reporting the highest ranked parse. On the other hand, even in the highest ranked analysis, relations that have a high probability score are more likely to be correct. Schneider et al. (2004) reports such a correlation: for example, PP-relations that are scored at least 60% likely have a precision 1-3% higher than those scored at least 30% likely.

The question is whether probability scores can be systematically exploited in the description of the New Englishes. The most ambiguous relations are generally the PP-attachment relations. Thus, if unusual verb-PP or noun-PP combinations are used in an L2-variety of English, this may show up in the probability scores. In a first experiment, we investigated the average probability per PP-attachment relation, *modpp* for noun-attachment and *pobj* for verb-attachment. The average probability for *modpp* is 82.8% in ICE-FIJI, and 83.1% in ICE-GB. The average probability for *pobj* is 79.9% in ICE-FIJI, and 78.7% in ICE-GB. There does not seem to be any significant difference. Furthermore, there is almost no genre variation: For *modpp*, all ICE-Fiji values are between 82.0% (press) and 83.5% (academic). The first experiment thus led to no conclusive results.

As a second experiment, we investigated at which level of informedness (backoff level) the decisions were taken. The use of unusual verb-PP or noun-PP combinations in an L2 variety may show up in the backoff level. Before we look at the results of this experiment, let us first explain what a backoff level is.

The statistical model used by the parser resolves parsing ambiguities based on lexical preference. Lexical preference is also called co-selection (Sinclair 2004) or native-like selection (see Pawley and Syder 1983 among others). It says e.g. that the noun *tea* is more likely to be modified by the adjective *strong* than by the adjective *powerful* (although these two adjectives are often synonymous), that *take* is more likely to be modified by the PP *into consideration* than by the PP *into respect*, etc. Knowledge about lexical preferences is obtained from the training corpus, i.e. the Penn Treebank, and used in the following way. Assume we are parsing the fragments

15) … ate chips with a fork

16) … ate chips with a steak

The PPs in both sentences can be attached to the verb *ate* or to the noun *chips*. Humans can resolve this ambiguity easily by world knowledge or lexical preferences, both of which are reflected in corpus frequencies: in a sufficiently large corpus, there are more instances where *with a fork* attaches to *eat* than to *chips*, and there are more instances where *with a steak* attaches to *chips* than to *eat.* Thus, if the parser chooses the more frequent attachment, it usually returns the correct syntactic analysis. The syntactically analysed corpora that are available today often do not contain enough information: no instance of PP-attachment containing the given governor verb (*eat*), preposition (*with*) and the noun inside the PP (e.g. *fork*) may be present, the disambiguation is not possible based on so-called full lexicalization. What can we

do? In many cases, less complete information is also sufficient: often, knowledge about semantic classes, or counts of the verb together with the preposition, or even the preposition itself (e.g. the preposition *of* indicates noun-attachment in the vast majority of cases) produced correct analyses in the majority of cases. This method, backing off from full lexicalization to partial lexicalization to a default (if even the preposition is unseen, the PP is attached to the noun, because noun-attachment is a bit more frequent) has been introduced in Collins and Brooks (1995) and is used in our parser for the most relations.
The working assumptions for the purposes of this paper are that in an L2 variety that is quite different from the L1 training corpus
(a) instances where the attachment decision can be based on full lexicalization are fewer, and that
(b) instances where only partial information is available are more frequent.

Schneider et al. (2004) report that there is a very strong correlation between backoff level and precision. For the GREVAL evaluation corpus, the correlation between backoff level and precision for the PP-attachment relations (modpp=nounpp, pobj=verbpp) is shown in Figure 12. Highly informed levels are much more likely to be correct.

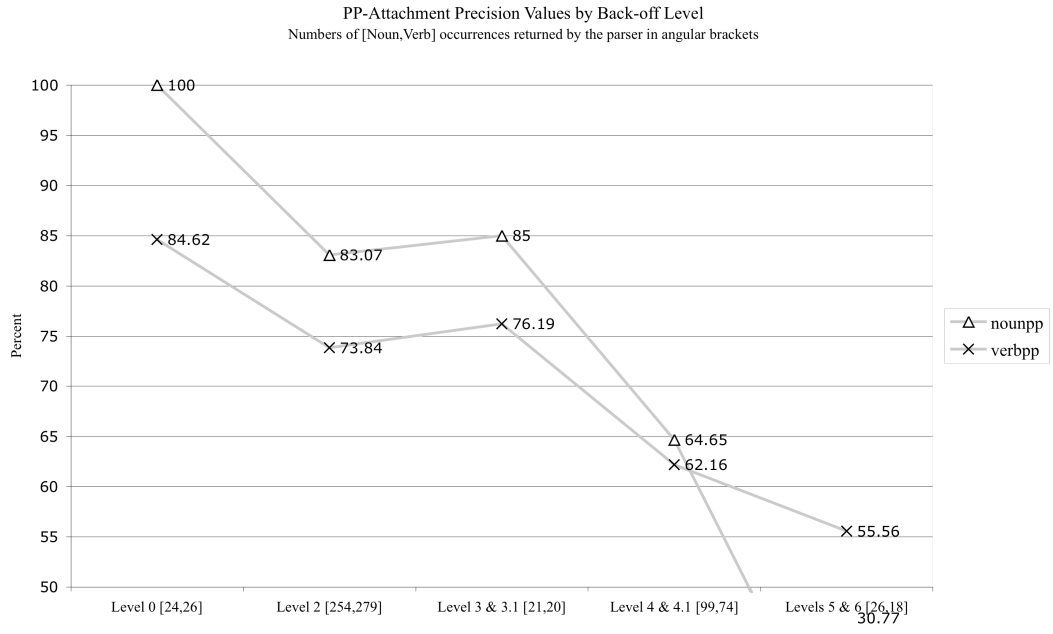

Figure 12: PP-Attachment precision by backoff level

Table 13 and Figure 13 show the backoff decision level percentages for ICE-FIJI, ICE-GB and ICE-INDIA for the PP-attachment relations (*modpp* and *pobj*). The semantic classes we have used are WordNet lexicographer files.

	FIJI%	GB%	India%	Description
0	12.02%	12.18%	11.07%	Governor+Prep+Noun (=full lexicalization)
2	57.34%	56.88%	57.08%	Governor+Prep

3	4.02%	3.52%	4.01%	GovernorClass+Prep+Noun
4	17.92%	17.04%	18.47%	GovernorClass+Prep+NounClass
5	1.71%	1.80%	1.92%	Prep+Noun
6	6.99%	8.57%	7.44%	Prep (=Preposition only)
7	0.00%	0.01%	0.00%	Default to noun-attachment

Table 13 Backoff Level Distributions

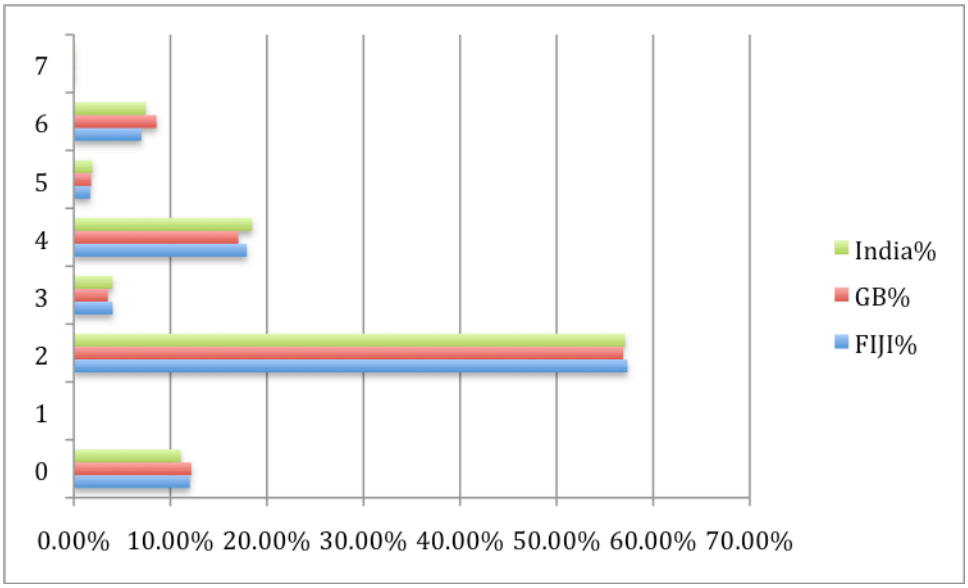

Figure 13: Backoff Level Distributions

The differences are very small: full lexicalization (level 0) is a bit more frequent in ICE-GB, most backoff levels except for preposition only (level 6) are a bit more frequent in ICE-FIJI and ICE-INDIA. The average backoff level for *modpp* is 2.63 in ICE-FIJI, and 2.70 in ICE-GB. The average backoff level for *pobj* is 2.55 in ICE-FIJI, and 2.71 in ICE-GB. There is almost no genre variation: For *modpp*, all values are between 2.53 (essay) and 2.66 (academic).

These figures indicate that from the parser's perspective, the disambiguation problem is very similar, and that the language varieties are similar. There is a strong relation between backoff level and parsing accuracy: the less informed the decision, the more likely it is to be incorrect.

6. Conclusion and outlook

The main finding of our pilot study is that the parser, which was originally designed to parse ENL varieties, also performs surprisingly well on corpora of ESL varieties. This means that heavily annotated (i.e. parsed) data can be usefully exploited in a partly corpus-driven approach to the description of New Englishes which we have called 'prong a'. This approach assumes that the parser generally annotates corpora of New Englishes with the same accuracy as it parses Native Englishes and that, as a consequence, the statistical differences in the use of constructions indicate differences between Native and ESL varieties of English. To fully exploit this approach, the results from the partly corpus-driven analyses will have to be supplemented by in-depth qualitative investigation. We will report on more detailed studies that make use of this approach in future papers.

Our second approach of using the parser as a heuristic tool assumes that nativized patterns are potential breaking points in the parsing process. The fact that parser performance is surprisingly robust on ESL varieties, as well, means that parser break-downs cannot, as initially

assumed, be used to uncover new features of these varieties. There is almost no direct correlation between parsing breakdowns and nativized constructions, both on a qualitative and a quantitative account.

References

Bautista, Maria Lourdes S. and Andrew B. Gonzales. 2006. "Southeast Asian Englishes." In Braj B. Kachru, Yamuna Kachru and Cecil L. Nelson (Eds), *The Handbook of World Englishes*. Malden MA: Blackwell, 130-44.

Carroll, John, Guido Minnen, and Edward Briscoe. 2003. "Parser evaluation: using a grammatical relation annotation scheme". In Anne Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora.* Dordrecht: Kluwer, 299–316.

Collins, Michael and James Brooks. 1995. "Prepositional Attachment through a Backed-off Model." In *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, MA.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Thesis, University of Pennsylvania, Philadelphia, PA.

Foley, Joseph A. (Ed.). 1988. *New Englishes: The Case of Singapore.* Singapore: Singapore University Press.

Hundt, Marianne. 1998. *New Zealand English Grammar. Fact or Fiction?* Amsterdam and Philadelphia: Benjamins.

Hundt, Marianne and Gerold Schneider. fc. "Article usage in Fiji English – An ethnic marker?" (in preparation).

Lehmann, Hans Martin and Gerold Schneider. 2009. "Parser-based analysis of syntax-lexis interactions." In Andreas Jucker, Daniel Schreier and Marianne Hundt (Eds). *Corpora: Pragmatics and Discourse*. (Proceedings of the 29th ICAME conference in Ascona, Switzerland, 14-18 May 2008). Amsterdam: Rodopi. pp. 477-502.

Mukherjee, Joybrato. 2005. *English Ditransitive Verbs: Aspects of Theory, Description, and a Usage-Based Model*. (Language and Computers). Amsterdam: Rodopi.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. "Describing verb-complementational profiles of New Englishes: a pilot study of Indian English." *English World-Wide* 27: 147-173.

Mukherjee, Joybrato and Stefan Gries. 2009. "Collostructional nativization in New Englishes: verb-construction associations in the International Corpus of English." *English World Wide* 30(1): 26-51.

Pawley, Andrew and Frances Syder. 1983. "Two Puzzles for linguistic theory: nativelike selection and nativelike fluency". In J. Richards and R. Schmidt (Eds.), *Language and Communication*. Longman.

Schneider, Edgar W. 2004. "How to trace structural nativization: Particle verbs in World Englishes." *World Englishes* 23: 227-49.

Schneider, Gerold, Fabio Rinaldi, Kaarel Kaljurand and Michael Hess, 2004. "Steps towards a GENIA Dependency Treebank". In *Proceedings of Treebanks and Linguistic Theories (TLT 2004)*, Tübingen, Germany.

Schneider, Gerold and Fabio Rinaldi and Kaarel Kaljurand and Michael Hess. 2005. "Closing the Gap: Cognitively Adequate, Fast Broad-Coverage Grammatical Role Parsing". *ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, Miami, FL.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis. Institute of Computational Linguistics, University of Zürich.

Schreier, Daniel 2003. *Isolation and Language Change: Sociohistorical and Contemporary Evidence from Tristan da Cunha English*. Houndsmills/Basingstoke and New York: Palgrave Macmillan.

Sand, Andrea. 1999. *Linguistic Variation in Jamaica. A Corpus-Based Study of Radio and Newspaper Usage*. Tübingen: Narr Francke Attempto.

Sedlatschek, Andreas. 2009. *Contemporary Indian English. Variation and Change.* Amsterdam and Philadelphia: Benjamins.

Seretan, Violeta and Eric Wehrli, 2006. "Accurate collocation extraction using a multilingual parser." In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney, Australia. Association for Computational Linguistics.

Sinclair, John. 2004. *Trust the text*. Language, corpus and discourse. Routledge.

Smadja, Frank, 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics, 19(1)*: 143–177.

---

[i] See for instance the homepage of the ICE project (http://ice-corpora.net/ice/) which reports that ICE-Philippines has been successfully tagged with the ICE tagger, but also that the tagging is still being checked. Hoffmann and Mukherjee (2006) make a use of tagged web-derived collections of text; in our paper, we use more richly annotated corpus material.

[ii] See http://www.ucl.ac.uk/english-usage/resources/grammar/index.htm

[iii] For Aarts and Close's project, see http://www.ucl.ac.uk/english-usage/projects/verb-phrase/index.htm.

[iv] One of the reasons why sentences in L1 corpora may fail to parse correctly are slips of the tongue/pen; for L2 varieties, the added complication is to distinguish similar phenomena from 'errors' and patterns that are examples of the nativization process, i.e. should be considered as 'features' of the new Englishes. The question of how to distinguish between errors, slips and features was discussed at a workshop at the ICAME conference in Lancaster in 2009.

[v] The parser is also designed to be reasonably fast, it parses the BNC in just over 24 hours on a fast server.

[vi] Note that these findings are based on sub-corpora rather than the complete one million word collections.

[vii] Note that punctuation symbols are counted as words.

[viii] There were two cases in a part of the corpus that is outside our current selection because the ethnicity of the author is not known.

[ix] Likewise, two of the *do* instances in ICE-NZ are parsing errors.